Cross-cultural and construct validity of the Animated Activity Questionnaire

Peter, Wilfred F.; de Vet, Henrika C.W.; Boers, Maarten; Harlaar, Jaap; Roorda, Leo D.; Poolman, Rudolf W.; Scholtes, Vanessa A.B.; Steultjens, Martijn; Hendry, Gordon J.; Roos, Ewa M.; Guillemin, Francis; Benedetti, Maria G.; Cavazzuti, Lorenzo; Escobar, Antonio; Dafinrud, Hanne; Terwee, Caroline B.

Published in:
Arthritis Care & Research

DOI:
10.1002/acr.23127

Publication date:
2017

Document Version
Author accepted manuscript

Original Article

Arthritis Care & Research
DOI 10.1002/acr.23127

Cross-cultural and construct validity of the Animated Activity Questionnaire

Wilfred F Peter[1,2] PT PhD, Henrika CW de Vet[1] PhD, Maarten Boers[3] MD PhD, Jaap Harlaar[4] PhD, Leo D Roorda[2] MD PhD, Rudolf W Poolman[5] MD PhD, Vanessa AB Scholtes[5] PhD, Martijn Steultjens[7] PhD, Gordon J Hendry[7] PhD, Ewa M Roos[8] PT PhD, Francis Guillemin[9] MD PhD, Maria G Benedetti[10] PhD, Lorenzo Cavazzuti[10] PT, Antonio Escobar[11] PhD, Hanne Dagfinrud[12] PhD, Caroline B Terwee[1] PhD.

[1]*Department of Epidemiology and Biostatistics and the EMGO Institute for Health and Care research, VU University Medical Center Amsterdam, the Netherlands.*

[2]*Amsterdam Rehabilitation Research Center | Reade, Amsterdam, the Netherlands.*

[3]*Department of Epidemiology and Biostatistics, and Amsterdam Rheumatology and immunology Center, VU University Medical Center, Amsterdam, the Netherlands.*

[4]*Department of Rehabilitation Medicine and MOVE research institute Amsterdam, VU University Medical Center Amsterdam, the Netherlands.*

[5]*Department of Orthopedics, Joint Research, OLVG, Amsterdam, the Netherlands.*

[6]*Glasgow Caledonian University, School of Health and Life Sciences, Glasgow, United Kingdom.*

[7]*Institute of Sports Science and Clinical Biomechanics, University of Southern Denmark, Odense, Denmark.*

[8]*University of Lorraine, EA 4360 APEMAC, Inserm CIC-EC 1433, University Hospital, Nancy, France.*

[9] *Istituto Ortopedico Rizzoli, Physical Medicine and Rehabilitation Unit, Bologna, Italy.*

[10]*Basurto University Hospital, Bizkaia, Spain. Health Service Research Network on Chronic Diseases (REDISSEC)*

[11] *Diakonhjemmet Hospital, Oslo, Norway*

Received: May 12, 2016; Revised: Sep 06, 2016; Accepted: Oct 11, 2016

Corresponding Author

W.F. Peter, PT, PhD

VU University Medical Center

Department of Epidemiology and Biostatistics, F-building MedFac

P.O. box 7057

1007 MB Amsterdam

The Netherlands

Phone: +31 20 4449829

E-mail: w.peter@vumc.nl

Running head : Cross-cultural validity of the AAQ

Funding: European League Against Rheumatism (EULAR)

Disclosure: There are no conflicts of interest

Abstract

Objective: The Animated Activity Questionnaire (AAQ) assesses activity limitations in patients with hip/knee osteoarthritis (HKOA), and consisting video animations of which patients choose the animation that best matches their own performance. The AAQ has shown good validity and reliability. This study aims to evaluate cross-cultural and construct validity of the AAQ.

Methods Cross-cultural validity was assessed using ordinal logistic regression analysis to evaluate Differential Item Functioning (DIF) across 7 languages. Construct validity was assessed by testing correlations between the AAQ, and a Patient Reported Outcome Measure (PROM) and performance-based tests.

Results: Data of 1239 patients were available. Compared to Dutch (n=279), none of the 17 items showed DIF in English (n=202), French (n=193), 1 item showed uniform DIF in Spanish (n=99) and Norwegian (n=62), and 2 items showed uniform DIF in Danish (n=201). In all these languages, the occurrence of DIF did not influence the total score, which remained comparable with the original Dutch version. For Italian (n=203) versus Dutch however, 6 items showed uniform DIF, and 1 item showed non-uniform DIF, indicating some problems with the cross-cultural validity between these countries. With regard to construct validity, the correlations with PROM (0.74) and performance-based tests (0.36-0.68) were partly as expected (> 0.60).

Conclusion: The AAQ, an innovative tool to measure activity limitations that can be placed on the continuum between PROMs and performance-based tests showed a good overall cross-cultural validity, and seems to have great potential for international use in research and daily clinical practice in many European countries.

Significants and Innovations

- An innovative tool on the continuum between PROMs and performance based tests to assess activity limitations in patients with hip- and knee osteoarthritis is developed: the Animated Activity Questionnaire (AAQ),
- The AAQ showed good cross-cultural validity in several languages.

Introduction

A comprehensive assessment of limitations in daily activities is essential in the management of hip and knee osteoarthritis (HKOA) in order to monitor the clinical course and the recovery after rehabilitation, pharmacological treatment or surgical interventions. Common methods to assess activity limitations are Patient Reported Outcome Measures (PROMs) [1,2] and performance-based tests. [3,4] Both methods have advantages, but also disadvantages. PROMs are considered easy to implement, inexpensive and harmless for the patient. But they are also highly dependent on the perception and the reference frame of the patient. [5,6] PROMs also require reading ability in the language at issue. Moreover, PROM scores are influenced by a large number of personal factors (e.g. body mass, depression, self-efficacy, fatigue and pain). [7-9] Performance-based tests, on the other hand, quantify the capacity of the patient on how well he or she is able to perform specific tasks. But these tests may be considered cumbersome and require physical presence of the patient. [9] It is also stated that tests administered in the clinic, do not represent a real life situation and only capture a snapshot of reality. [10]

The Animated Activity Questionnaire (AAQ) was recently developed as a new method to assess activity limitations in patients with HKOA, for use in research and clinical practice. [11] The AAQ uses videos in which an animation of a basic daily activity is shown, and performed with different levels of difficulty. Patients are asked to choose the animation that best matches their own performance. The AAQ combines the advantages of self-report questionnaires and performance-based test. The AAQ is easy to implement, inexpensive, harmless for patients, and in addition no comprehensive language understanding, except for directions and internet navigating is required, which makes the AAQ accessible for people with low literacy and non-native speakers. In addition the AAQ does not need an intensive

forward and backward translation effort in case of international use as in questionnaires, and the presence of patients in a clinic is not required as in performance-based tests. The AAQ showed good content validity, construct validity and reliability.[11] Moreover, the AAQ appeared to measure activity limitations closely mimicking real life situations. [12]

To evaluate the suitability of the AAQ across countries cross-cultural validity should be studied. Cross-cultural validity has been defined as 'the degree to which the performance of items on a translated or culturally adapted instrument is an adequate reflection of the performance of items in the original version of the instrument' [13]. Cross-cultural validity can be assessed by Differential Item Functioning (DIF) analyses. Ideally, patients from different countries with the same level of activity limitations should have the same score on each AAQ item (no DIF). [13] We hypothesized that, with respect to the AAQ, DIF due to language differences will not be prevalent because minimal translation, except for directions, is needed. In earlier studies, construct validity of the AAQ was tested in a small group of patients from one country only.[11,12] These results should be confirmed in larger groups of patients. Therefore the aim of our study was to study cross-cultural and construct validity of the AAQ in larger groups of patients from 7 European countries.

Patients and Methods

Participants

This study was conducted in 7 European countries; Denmark, France, Italy, the Netherlands, Norway, Spain, and United Kingdom with all different languages. In all participating countries patients aged over 18 years, and with a diagnosis of hip and/or knee OA according the ACR criteria [14,15], were invited by phone or when they visited the clinic where they receive treatment, to participate in the study. If they were willing to participate, an

information leaflet, an informed consent form, and a pre-stamped, pre-addressed envelope were sent or given to them personally. After receiving the signed informed consent, the patients were included in the study. A consecutive sample of patients was recruited from different health care setting such as primary care, in-patient rehabilitation, and hospitals. The goal was to include 200 patients in each country since this is considered an adequate sample size for DIF analyses. [16]

Ethical approval

This study was conducted in accordance with the Handbook for Good Clinical Research Practice of the World Health Organization, and Declaration of Helsinki principles [http://www.wma.net/en/30publications/10policies/b3/] and, if required, approved by the corresponding Medical Ethics Committees of the participating countries.

Assessments

The participants were sent an e-mail including an URL to an online questionnaire. They were asked to complete the AAQ which contains videos of 17 basic daily life activities [11]: ascending and descending stairs (2 items); walking outside on a flat surface (1 item); walking outside on uneven terrain (1 item); walking inside (1 item); ascending and descending a slope (2 items); picking up an object from the floor (1 item); rising from sitting on the floor (1 item); rising and sitting down from a chair, a sofa and a toilet (6 items); and taking off and putting on shoes (2 items). Of each activity three to five videos were shown with an increase of difficulty of performance, resulting in 3-5 response options. All videos of an activity were shown simultaneously, ordered by level of difficulty on the screen, to facilitate comparison of performance and level of difficulty. The first video of each activity represents optimal performance, and the last video represents the highest level of difficulty in performance. The videos could be played as many times as the patient wanted to see them. All participants were

instructed as follows: "the video that best matches my own performance is ...". Each activity also offer the response option 'unable to perform'. Collaborating research partners in the participating countries translated this question and the scoring instructions into their own language.

Each activity was scored on a scale from 0 to 3, 4 or 5 depending on the number of response videos for each activity. A normalized score from 0-100 was calculated for each activity. The total score was calculated by taking the mean scores of the normalized scores of all activities and dividing that by 17, with higher scores indicate less activity limitations. Two examples of an item, and how the AAQ is developed, are available at: http://www.kmin-vumc.nl/_16_0.html

Directly after completing the AAQ, patients were asked to complete a Patient Reported Outcome Measure (PROM); the 'Function, daily living'(ADL) subscale of the HOOS [1] (i.e. Hip disability and Osteoarthritis Outcome Score) or KOOS [2] (i.e. Knee injury and Osteoarthritis Outcome Score) depending on the affected joint. The H/KOOS ADL subscale contains seventeen items assessing perceived limitations in physical functioning. Each item was rated on a five-point rating scale (i.e. 0-4). Scores were transformed to a 0-100 score with higher scores corresponding to less activity limitations. The HOOS and KOOS showed adequate content and construct validity in most of the participating countries. [1,2, 17-20]

In order to prevent contamination we changed the order of the AAQ and H/KOOS ADL subscale in the second half of the patients in each country.

Finally, in each country patients were asked to execute three performance-based tests in the clinical setting of the participating hospital or outpatient clinic in the following predetermined order: the Stair Climbing Test [9, 21], and the Timed Up and Go test [22, 23], both measuring the time in which the activity is performed, and the 30 second Chair Stands Test.

[24, 25] which takes to number of sit to stands that was performed within 30 seconds. The chosen performance-based tests were chosen from the most feasible, reliable and responsive measures recommended by OsteoArthritis Research Society International.[26]

Patients were asked in consecutive order to execute the performance-based tests until a minimum number of 35 patients per country participated.

Statistical analyses

Descriptive statistics were used to describe the study population with regard to age, gender, Body Mass Index (BMI), affected joint (i.e. knee, hip or both), total joint replacement surgery (none, unilateral, bilateral), physical therapy treatment, the mean AAQ score, mean H/KOOS subscale score, means of the performance-based tests, and pain.

First, we tested the AAQ for unidimensionality in each participating country and in the total patient group by means of a confirmatory factor analysis, as a prerequisite to perform the cross-cultural validity analysis. Model fit was evaluated by estimating the Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), and Tucker-Lewis fit Index (TLI). RMSEA close to 0.06 or lower, CFI close to 0.95 or higher, and TLI close to 0.95 or higher indicate good model fit. [27]

In addition, for cross-cultural validity an ordinal logistic regression analysis as described by Petersen et al. [28] was used to assess Differential Item Functioning (DIF) across countries. As the AAQ was developed in the Netherlands, the Dutch version of the AAQ was considered the original version. DIF was assessed for each AAQ item, and for each country separately. In the ordinal regression analyses the dependent variable was the AAQ item score, and the independent variables were the group variable (two groups per analysis, with the Dutch version as the reference group), the total AAQ score, and the interaction term between the group variable and the total AAQ score. First we tested for non-uniform DIF. A pseudo R-

square change score according Nagelkerke with a magnitude larger than 0.035 and a significant interaction term (p-value < 0.001) [29] between the AAQ total score and the group variable was considered as non-uniform DIF. If there was no non-uniform DIF, we tested for uniform DIF. An odds ratio (OR) of the group variable with a magnitude outside the interval 0.53–1.89 and a statistical significance with a p-value < 0.001 was used as a criterion for uniform DIF. [30] An OR below 0.53 indicates that a patient from the country under study, with a similar functional level as a Dutch patient, score lower on the item. So, the execution of the activity by a patient from the country at issue seems to be more difficult than for a Dutch patient. If the OR is above 1.89 this indicates that a patient from the other country scores higher, and executing the activity is less difficult than for the Dutch patients. Ordinal logistic regression analyses were adjusted for gender, age, height, weight, affected joint, and presence of a hip or knee prosthesis. If non-uniform DIF was found for an item, the scores of that item were visualised by plotting the probability of a response option (in this case choosing a certain video) against the total AAQ score for both countries, to show how much the responses of the two countries are different and in which direction.

The impact of item(s) with DIF on the total score was determined by calculating the correlation between the AAQ score with and without the DIF item(s). A correlation higher than 0.95 was considered as no important impact of DIF on the total AAQ score.

Finally, to assess construct validity, correlations were calculated between the AAQ score and the H/KOOS ADL subscale by means of the correlation coefficients (Spearman's- or Pearson's correlation coefficient, depending on distribution of data). We calculated the average correlation of the AAQ score and the three performance-based tests, and the correlations between the AAQ and each performance-based test separately. The average score was used because of the overlap with these tests with activities in the AAQ as well in the H/KOOS ADL subscale, who both comprise a combination of different activities. We

hypothesized that all correlation coefficients were > 0.60. In order to average the correlations between AAQ score and the performance-based tests, scores of the three different performance-based tests (i.e. Stair Climbing Test, Timed Up and Go test, and the 30 sec. Chair Stands Test) were transformed into Fisher's Z scores. After summarizing and averaging the three scores, the score was transferred back into a correlation coefficient.

Data were analysed with the SPSS statistical package (version 20.0, SPSS, Chicago, Illinois), and Mplus version 6.11.

Results:

Data of 1239 patients were available from: Denmark (N=201), France (N=193), Italy (N=203), the Netherlands (N=279), Norway (N=62), Spain (N=99), and the United Kingdom (N=202).

Patient characteristics

Table 1 shows that the mean age of the patients was 64.2, 72% were female, and the mean BMI was 28. Of participating patients 58% had knee problems, 26% had hip problems, while in 16% of the patients both joints were affected. In 35% of the patients unilateral, in 2% bilateral, 18% knee, and 22% hip joint replacement surgery had taken place, and 45% received physical therapy treatment at the moment the assessments took place. The mean AAQ score was 77 (SD 18.5), and the mean H/KOOS ADL subscale score 66 (SD 20.5). The means of the Timed Up and Go test, Stair Climbing Test and 30 sec. Chair Stands Test were 11.3 seconds, 17.5 seconds, and 10.5 sit to stands, respectively.

Confirmatory Factor Analyses showed good model fit for unidimensionality of the AAQ in the total patient group for two out of three estimations (CFI = 0.957, TLI = 0.951, and RMSEA = 0.144 (90% confidence Interval 0.139-0.148). Individual country versions showed similar results (Table 2).

Results of DIF analyses are shown in Table 3. Compared to the Dutch version, none of the 17 items showed DIF in English or French patients. Uniform DIF was found in one item for Spanish versus Dutch (walking inside; OR 0.29), and one item for Norwegian versus Dutch (walking inside; OR 0.16). Walking indoors was more difficult for Spanish or Norwegian patients respectively compared to Dutch patients. For Danish versus Dutch versions of AAQ, uniform DIF was found in two items (walking outside on uneven terrain; OR 0.45, walking inside; OR 0.43), representing more difficulty in executing the activities for Danish patients compared to Dutch patients. An example of uniform DIF is shown in Figure 1. For Italian versus Dutch 6 items showed uniform DIF (descending stairs; OR 0.35, walking outside on uneven terrain; OR 0.34, picking an object from the floor; OR 0.26, and rising from the floor; OR 0.26), representing more difficulty in executing all four activities for Italian patients compared to Dutch patients. Rising from a sofa; OR 10.40, and sitting down on a sofa; OR 3.75, were less difficult for Italian patients compared to Dutch patients. One item in the Italian AAQ version showed non-uniform DIF (rising from a toilet; Nagelkerke 0.06). For patients with high limitations in activities, Italian patients have less difficulty with rising from a toilet compared to Dutch patients. However, for patients with less limitation in activities the difficulty of the item is more similar between Italian and Dutch patients. (see Figure 1)

There was no important influence of uniform DIF on the total AAQ score for Spain and Norway; Spearman's correlation between AAQ score with and without DIF item were 0.98 and 0.99, respectively. For Denmark the Spearman's correlations between AAQ score with and without the two DIF items was 0.99. For Italy the Spearman's correlations between AAQ score with and without the seven DIF items was 0.98 (data not shown).

Regarding construct validity, Table 4 shows that in the total group of 1239 patients the total AAQ score correlated highly (0.74) with the total H/KOOS ADL subscale. In a subgroup

of 272 patients, in which the performance-based tests were executed, the correlation of the total AAQ score with scores on the performance-based tests was lower than expected (0.55). For the individual performance-based tests the correlations with the AAQ were 0.68 (Stair Climbing Test), 0.59 (Timed Up and Go test), and 0.36 (30 sec. Chair Stands Test).
In Table 5 the results of the correlations within each country are shown, with a high correlation (>0.60) between AAQ and H/KOOS ADL subscale for each country, ranging from 0.64 to 0.85. The correlations between AAQ and performance-based tests range from 0.48 to 0.70, with a moderate correlation for Denmark (0.52), Italy (0.48) and borderline high correlation for France (0.60), while the correlations for The Netherlands, Spain and United Kingdom were high (0.68, 0.70 and 0.67, respectively). In Norway no performance tests data were collected.

Discussion

In this study the AAQ was shown to have good cross-cultural validity, but there were unexpected findings in construct validation. Regarding the former, some cultural differences were seen between Dutch and Italian versions (DIF in 7 items), but the impact of DIF on the total AAQ score was negligible.

Regarding construct validity, AAQ appeared to be unidimensional, although the RMSEA indices (0.144) did not reach the level of good fit (<0.06). When the AAQ was developed we expected that the AAQ would correlate higher with performance-based tests than with other self-report questionnaires because the AAQ resembles more the real life situation. However, all validation studies showed the opposite results: AAQ showed a higher correlation with the H/KOOS ADL subscale (0.74) than hypothesized, and similar to that found previously (i.e. 0.76 [11] and 0.79 [12]). Also, the correlation between AAQ and performance-based tests

(0.55) was lower than expected, and also lower than in previous studies (i.e. 0.62 [11] and 0.73 [12]).

A possible explanation is that the AAQ comprises more activities than the three activities of the performance-based tests, such as picking up an object from the floor, rising from the floor, and putting on and taking off shoes. On the other hand the H/KOOS ADL subscale also comprises other activities, such as going shopping or sitting, and the correlation with AAQ is high (0.74). We favor a second explanation: the AAQ is a Patient Reported Outcome Measure (PROM) that measures a new construct, closely related to perception of performance and actual life performance. Apparently, the perception of the patient in completing the AAQ is playing a bigger role than we assumed, and while performance-based tests measure 'just' capacity in the 'lab', the AAQ goes beyond this and refers to an actual life performance, for which indications are shown in earlier research in which the AAQ was highly correlated (0.83) with home-recorded videos of activities performed by the patients at their own home.[12]

Regarding cross-cultural validity this study showed that the AAQ has potential to yield comparable scores across countries with different languages. The total score showed minimal DIF and thus seems to be comparable across the countries under study. However, scores on individual items, useful for clinicians in daily practice, cannot always be compared across countries because of DIF. The advantage of the AAQ over other self-report questionnaires is that no translation, except for directions, is needed, and that the AAQ is easier to use for people with low levels of literacy.

DIF between countries is usually caused by differences in translation and culture. For the AAQ, no intensive forward and backward translation is needed. Since the AAQ is a computerized animation questionnaire which shows videos and no written questions (for

which translation of items and response options is necessary), it is expected that the observed DIF between countries is mainly caused by cultural differences and not language issues.

Although DIF occurred in 7 out of 17 items in Italian version, the impact on the total score was negligible, similar as in other languages in the study, since the difference in AAQ scores with and without DIF items was small with a correlation of 0.98. This is probably caused by the fact that items with an odds ratio (OR) score beneath one (4 items with ORs of 0.35, 0.34, 0.26, and 0.26, respectively) were neutralized by the effect of items with an odds ratio above 1 (2 items with odds ratios of 10.40 and 3.75 respectively), and the item with non-uniform DIF.

Nevertheless, more in-depth analyses are necessary to understand the occurrence of DIF in Italy; Italian patients showed more difficulty in executing descending stairs, walking on uneven surface, picking an object from floor and rising from the floor and less difficulty in executing sitting down and rising from a sofa compared to Dutch patients with a similar level of activity limitations. Moreover, a part of the Italian patients reported to have more difficulty in rising from a toilet compared to Dutch patients. Differences between the populations in weight and height were adjusted for. Although adjusted for the presence of a hip or knee prosthesis, a substantial higher proportions of Italian patients (66%) had joint replacement surgery compared to the Dutch patients (38%) which can play a role in country differences in severity of the disease and activity limitations due to joint mobility differences, which was not assessed. Also different national standards in stairs, or sofa heights may be an explanation for country differences.

A qualitative approach should be added to the statistical DIF analyses for the interpretation of DIF since a low agreement has been found in the past between expert review of items and statistical analyses.[30, 31] With qualitative methods more socio-psychological explanations can be explored for underlying reasons for DIF. Collins [32] mentioned the social-cognitive

theory of survey response which can play a role in answering questions on a questionnaire. The theory involves 4 sequential cognitive tasks in answering questions: (1) question interpretation, (2) retrieval of information from memory to answer the question, (3) judgment formation and response formation, and (4) response evaluation and response editing. In each of the tasks, differences between countries can occur, based on personal and contextual factors. This theory is focused on written questionnaires. With the AAQ also other cognitive processes will play a role, such as self-reflection of movements. How well can a patient self-reflect on stair climbing or rising from the floor when choosing a corresponding performance level of the activity shown in videos as in the AAQ? Differences between Italy and the Netherlands might be based on more optimistic or pessimistic attitude in responding in one of the two countries or differences in response styles (i.e. patients in one country tends to select the end points of the scale more often). [33]

A methodological consideration with regard to our study is that we used the pseudo R-square change according Nagelkerke of 0.035 as a cut-off point for non-uniform DIF. [29] The Zumbo-Thomas procedure [16] uses a far more conservative cut-off point for DIF of 0.13 which would have resulted in less DIF in our study. On the other hand there are also studies that used a cut-off point of 0.02. [34, 35]

Another point for discussion is the variable number of response options of the items of the AAQ, ranging between 4 and 6. In theory this difference could weaken the unidimensionality of the scale, as items with the same number of response options tend to correlate higher with each other than items with a different number of response options. But in our data items with the same response options were not systematically correlated higher with each other than items with different response options.

A limitation of our study is that the results of the cross-cultural validity analyses for Spain (n=99) and Norway (n=62) should be interpreted with caution because the recommended 200 respondents for adequate DIF analyses [16] was not achieved. Another limitation is having the instructions i.e. "the video that best matches my own performance is", instead of including the temporal factor such as "my own performance today/in the previous week/in the previous month". On a paper based PROM participants can turn back over and re-read instructions which usually say something like "thinking about your ability today/in the past week etc.

In conclusion, the AAQ showed a good overall cross-cultural validity, and presents an innovative way of measuring a new construct which is self-reported, and that can be placed on the continuum between PROMs and performance-based tests. The AAQ seems to have great potential for international use in research and daily clinical practice. An online version to be used by patients, clinicians and researchers is in preparation. To get more insight in the construct of the AAQ, future research of the AAQ should focus on qualitative research which is necessary to explore explanations for DIF in Italy, and more data should be collected in Italy to confirm or refute the results in this study.

Acknowledgements

The following colleagues and health care institutes are acknowledged for collecting the data in the different European countries: Anne Marie Rosager at SANO Rehabilitation Center in Skælskør (Denmark), Amandine Vallata, Isabelle Petitgenet at Inserm CIC-EC 1433, University hospital, Nancy (France), Lorenzo Cavazzuti at Istituto Ortopedico Rizzoli, Bologna, (Italy), Turid Høysveen at Ullernklinikken Manual therapy and Rehab Oslo (Norway), Kim Brown, Emma Mcloughlin, Anna Thornhill at Solent NHS Trust (United Kingdom), Natalia Andrea Rivera Garcia at Basurto University Hospital, Bizkaia (Spain), and Reade, centre for rehabilitation and rheumatology, and Joint Research, OLVG Hospital,

Amsterdam (the Netherlands), for participating in the study.

References

1. Groot IB de, Reijman M, Terwee CB et al. Validation of the Dutch version of the Hip disability and Osteoarthritis Outcome Score. Osteoarthritis Cartilage 2007; 15: 104-9
2. Groot IB de, Favejee MM, Reijman M, Verhaar JA, Terwee CB. The Dutch version of the Knee Injury and Osteoarthritis Outcome Score: a validation study. Health Qual Life Outcomes 2008; 6 :16.
3. Bennel K, Dobson F, Hinman R. Measures of Physical Performance Assessments; Self-Paced Walk Test (SPWT), Stair Climb Test (SCT), Six-Minute Walk Test (6MWT), Chair Stand Test (CST), Timed Up & Go (TUG), Sock Test, Lift and Carry Test (LCT), and Car Task. Arthritis Care & Research 2011; 11: 350–70.
4. Dobson F, Hinman RS, Roos EM, et al. OARSI recommended performance-based tests to assess physical function in people diagnosed with hip or knee osteoarthritis, Osteoarthritis and Cartilage 2013; 21: 1042-52.
5. Stratford PW, Kennedy DM. Performance measures were necessary to obtain a complete picture of osteoarthritic patients. J Clin Epidemiol 2006; 59: 160-167.
6. Fayers PM, Langston AL, Robertson C. Implicit self-comparisons against others could bias quality of life assessments. J Clin Epidemiol 2007; 60: 1034-9.
7. Maly MR, Cosigan PA, Olney SJ. Determinants of Self-Report Outcome Measures in People With Knee Osteoarthritis. Arch Phys Med Rehabil 2006; 87: 96-104,.
8. Terwee CB, Slikke RMA van der, Lummel RC van, Bennink JB, Meijers WGH, Vet HCW de. Self-reported Physical Functioning was more Influenced by Pain than Performance-Based Physical Functioning in Knee-Osteoarthritis Patients. J Clin Epidemiol 2006; 59 : 724-731.

9. Kennedy D, Stratford PW, Pagura SM, Walsh M, Woodhouse LJ. Comparison of gender and group differences in self-report and physical performance measures in total hip and knee Arthroplasty 2002; 17: 70-7

10. Wittink H, Rogers W, Sukiennik A, Carr DB. Physicial functioning: self-report and performance measures are related but distinct. Spine 2003; 28: 2407-13.

11. Peter WF, Loos M, de Vet HC, Boers M, Harlaar J, Roorda LD, Poolman RW, Scholtes VA, Boogaard J, Buitelaar H, Steultjens M, Roos EM, Guillemin F, Rat AC, Benedetti MG, Escobar A, Østerås N, Terwee CB. Development and preliminary testing of a computerized animated activity questionnaire in patients with hip and knee osteoarthritis. Arthritis Care Res (Hoboken) 2015; 67: 32-9.

12. Peter WF, Loos M, van den Hoek J, Terwee CB. Validation of the Animated Activity Questionnaire (AAQ) for patients with hip and knee osteoarthritis: comparison to home-recorded videos. Rheumatol Int 2015; 35: 1399-408.

13. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemio. 2010; 63: 737-45.

14. Altman R, Alarcon G, Appelrouth D et al. The American College of Rheumatology criteria for the classification and reporting of osteoarthritis of the hip. Arthritis Rheum 1991; 34: 505-14.

15. Altman R, Asch E, Bloch D et al. Development of criteria for the classification and reporting of osteoarthritis. Classification of osteoarthritis of the knee. Diagnostic and Therapeutic Criteria Committee of the American Rheumatism Association. Arthritis Rheum 1986; 29: 1039-49.

16. Zumbo BD: A handbook on the theory and methods of differential item

functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Ottowa, ON: Directorate of Human Research and Evaluation, Department of National Defense 1999.

17. Collins NJ, Prinsen CA, Christensen R, Bartels EM, Terwee CB, Roos EM. Knee Injury and Osteoarthritis Outcome Score (KOOS): systematic review and meta-analysis of measurement properties. Osteoarthritis Cartilage. 2016; pii: S1063-4584(16)01071-2.
18. Collins NJ, Roos EM. Patient-reported outcomes for total hip and knee arthroplasty: commonly used instruments and attributes of a "good" measure. Clin Geriatr Med 2012; 28: 367-94.
19. Ornetti P, Parratte S, Gossec L, Tavernier C, Argenson JN, Roos EM, Guillemin F, Maillefert JF. Cross-cultural adaptation and validation of the French version of the Hip disability and Osteoarthritis Outcome Score (HOOS) in hip osteoarthritis patients. Osteoarthritis Cartilage 2010; 18: 522-9.
20. Nilsdotter AK, Lohmander LS, Klässbo M, Roos EM. Hip disability and osteoarthritis outcome score (HOOS)--validity and responsiveness in total hip replacement. BMC Musculoskelet Disord 2003; 4: 10.
21. Rejeski WJ, Ettinger WH Jr, Schumaker S, James P, Burns R, Elam JT. Assessing performance-related disability in patients with knee osteoarthritis. Osteoarthritis Cartilage 1995; 3: 157-67.
22. Stratford PW, Kennedy DM, Woodhouse LJ. Performance measures provide assessments of pain and function in people with advanced osteoarthritis of the hip or knee. Phys Ther 2006; 86: 1489-96.
23. Steffen TM, Hacker TA, Mollinger L. Age- and gender-related test performance in community-dwelling elderly people: Six-Minute Walk Test, Berg Balance Scale, Timed Up & Go Test, and gait speeds. Phys Ther 2002; 82: 128-37.

24. Gill S, McBurney H. Reliability of performance-based measures in people awaiting joint replacement surgery of the hip or knee. Physiother Res Int 2008; 13: 141e52.

25. Jones CJ, Rikli RE, Beam WC. A 30-s chair-stand test as a measure of lower body strength in community-residing older adults. Res Q Exerc Sport 1999; 70: 113e9.

26. Dobson F, Hinman RS, Roos EM, Abbott JH, Stratford P, Davis AM, Buchbinder R, Snyder-Mackler L, Henrotin Y, Thumboo J, Hansen P, Bennell KL. OARSI recommended performance-based tests to assess physical function in people diagnosed with hip or knee osteoarthritis. Osteoarthritis Cartilage 2013; 21: 1042-52.

27. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modelling 1999; 6: 1-55.

28. Petersen MA, Groenvold M, Bjorner JB, Aaronson N, Conroy T, Cull A, Fayers P, Hjermstad M, Sprangers M, Sullivan M; European Organisation for Research and Treatment of Cancer Quality of Life Group. Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire. Qual Life Res 2003; 12: 373-85.

29. Jodoin MG, Gierl MJ, Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. Applied Measurement in Education 2001; 14: 329-349.

30. Benson J, Hutchinson SR: The state of the art in bias research in the United States. European Review of Applied Psychology 1997; 47: 281-294.

31. Roussos L, Stout W: A multidimensionality-based DIF analysis paradigm. Applied Psychological Measurement 1996; 20: 355-371.

32. Collins D. Pretesting survey instruments: an overview of cognitive methods. Quality of Life Research 2003; 12: 229-238.

33. Chen C, Lee S, Stevenson H. Response style and cross-cultural comparisons of rating scales among East Asians and North American students. Psychol Sci 1995; 6: 170-175

34. Bjorner JB, Kosinski M, Ware JE: Calibration of an item pool for assessing the burden of headaches: An application of item response theory to the headache impact test (HIT). Quality of Life Research 2003; 12: 913-933.

35. Schmidt S, Mühlan H, Power M: The EUROHIS-QOL 8-item index: Psychometric results of a cross-cultural field study. European Journal of Public Health Advance Access 2006; 16: 420-428.

Table 1. Patient Characteristics

	Total group (n=1239)	Denmark (n=201)	France (n=193)	Italy (n=203)	the Netherlands (n=279)	Norway (n=62)	Spain (n=99)	United Kingdom (n=202)	Subgroup validity performance-based tests (n=272, all countries)
Sex, female (%)	887 (71.6%)	164 (81.6%)	126 (65.3%)	146 (71.9%)	193 (69.2%)	54 (87.1%)	80 (80.8%)	124 (61.4%)	201 (73.9%)
Age, mean (SD)	64.2 (9.9)	62.7 (10.4)	65.8 (7.2)	63.9 (12.0)	65.3 (7.9)	67.8 (10.0)	67.4 (10.1)	60.3 (10.6)	63.5 (8.5)
BMI, mean (SD)	27.7 (5.5)	29.0 (5.9)	28.5 (5.8)	23.7 (6.0)	29.0 (6.0)	25.7 (4.0)	28.9 (5.0)	28.9 (5.0)	30.0 (6.4)
Joint affected									
knee	719 (58.0%)	117 (58.2%)	117 (60.6%)	97 (47.8%)	156 (55.9%)	30 (48.4%)	77 (77.8%)	125 (61.9%)	163 (59.9%)
hip	324 (26.2%)	52 (25.9%)	49 (25.4%)	106 (52.2%)	48 (17.2%)	20 (32.3%)	20 (20.2%)	29 (14.4%)	64 (23.6%)
knee and hip	196 (15.8%)	32 (15.9%)	27 (14.0%)	0	75 (26.9%)	12 (19.4%)	2 (2.0%)	48 (23.8%)	45 (16.5%)
Total joint replacement									
none	768 (62.0%)	135 (67.2%)	151 (78.2%)	70 (34.5%)	166 (59.5%)	33 (53.2%)	88 (88.9%)	125 (61.9%)	191 (70.2%)
knee	222 (17.9%)	35 (17.4%)	21 (10.9%)	70 (34.5%)	52 (18.6%)	19 (30.6%)	3 (3.0%)	22 (10.9%)	34 (12.5%)
hip	278 (22.4%)	40 (19.9%)	25 (13.0%)	63 (31.0%)	68 (24.4%)	12 (19.4%)	9 (9.1%)	61 (30.2%)	50 (18.4%)
unilateral	433 (34.9%)	57 (28.4%)	38 (19.7%)	133 (65.5%)	106 (38.0%)	27 (43.5%)	10 (10.1%)	71 (35.1%)	78 (28.7%)
bilateral	29 (2.3%)	9 (4.5%)	4 (2.1%)	0	7 (2.5%)	2 (3.2%)	1 (1.0%)	6 (3.0%)	3 (1.1%)
Timed Up and Go test, mean (sec.)	11.3 (4.2)	9.6 (3.9)	9.9 (2.6)	15.0 (2.8)	11.3 (5.0)	*	11.9 (4.6)	11.3 (3.4)	11.7 (4.2)
Stair Climbing Test, mean (sec.)	17.5 (8.6)	15.7 (7.1)	14.7 (6.6)	27.4 (5.4)	18.2 (10.1)	*	15.2 (5.8)	16.3 (8.3)	18.5 (8.8)
30 sec. Chair Stand Test, mean (counts)	10.5 (7.1)	10.6 (4.2)	10.5 (3.7)	13.5 (2.4)	10.4 (12.2)	*	9.1 (4.4)	9.5 (4.1)	10.7 (6.8)
Physical therapy treatment? Yes	554 (44.7%)	181 (90.0%)	14 (7.3%)	162 (79.8%)	89 (31.9%)	57 (91.9%)	4 (4.0%)	47 (23.3%)	102 (37.5%)
AAQ score (0-100), mean (SD)	76.9 (18.5)	76.6 (15.6)	86.7 (14.7)	64.9 (19.2)	77.6 (17.8)	79.3 (15.6)	80.3 (18.4)	76.8 (19.3)	77.1 (17.7)
H/KOOS ADL subscale (0-100), mean (SD)	66.4 (20.5)	62.4 (16.5)	73.4 (19.9)	69.6 (17.6)	64.2 (21.7)	61.9 (19.8)	61.7 (24.1)	67.3 (22.1)	64.8 (20.1)
Pain (0-10), mean (SD)	4.3 (2.6)	5.4 (2.4)	3.4 (2.3)	3.8 (2.6)	4.5 (2.7)	4.2 (2.2)	4.7 (2.6)	4.3 (2.6)	4.7 (2.6)

Table 2 Confirmatory Factor Analyses of the AAQ in 7 different European countries

		CFI	TLI	RMSEA (90% Confidence Interval)
Denmark	n=201	0.913	0.901	0.166 (0.155-0.178)
France	n=193	0.973	0.969	0.111 (0.099-0.123)
Italy	n=203	0.982	0.980	0.110 (0.099-0.122)
The Netherlands	n=279	0.963	0.957	0.127 (0.117-0.136)
Norway	n=62	0.934	0.925	0.147 (0.124-0.169)
Spain	n=99	0.980	0.977	0.098 (0.079-0.116)
United Kingdom	n=202	0.982	0.979	0.113 (0.102-0.125)
Total	n=1239	0.957	0.951	0.144 (0.139-0.148)

CFI = Comparative Fit Index
TLI = Tucker-Lewis index
RMSEA = Root Mean Square Error of Approximation

Table 3 Odds Ratio's and Pseudo R-square values for Uniform Differential Item Functioning (DIF), and Non-Uniform DIF respectively, in 17 items of the Animated Activity Questionnaire (AAQ)

	Odds Ratio	95% Confidence interval	p-value	Pseudo R-square change (Nagelkerke) #	P-value #
1. Ascending stairs					
Italy (n=203)	0.42	0.24-0.74	0.002	0.006	0.74
United Kingdom (n=202)	1.30	0.82-2.07	0.26	0.001	0.006
Denmark (n=201)	1.11	0.71-1.63	0.62	0.003	0.14
France (n=193)	1.52	0.80-2.88	0.20	0.004	0.39
Spain (n=99)	0.39	0.18-0.75	0.005	0.012	0.46
Norway (n=62)	1.31	0.41-1.52	0.43	0.001	0.49
2. Descending stairs					
Italy (n=203)	0.35*	0.27-0.63	<0.001*	0.018	0.12
United Kingdom (n=202)	0.86	0.58-1.26	0.44	0.002	0.24
Denmark (n=201)	1.22	0.83-1.76	0.20	0.002	0.94
France (n=193)	2.08	1.19-3.65	0.01	0.018	0.03
Spain (n=99)	1.13	0.61-2.15	0.70	0.000	0.82
Norway (n=62)	1.10	0.68-2.20	0.75	0.001	0.40
3. Walking outside on a flat surface					
Italy (n=203)	0.86	0.43-1.13	0.61	0.012	0.32
United Kingdom (n=202)	1.11	0.71-1.73	0.65	0.000	0.40
Denmark (n=201)	0.51	0.33-0.77	0.001	0.012	0.32
France (n=193)	2.24	1.13-4.41	0.02	0.008	0.82
Spain (n=99)	1.77	0.79-4.12	0.13	0.007	0.09
Norway (n=62)	0.65	0.33-1.18	0.20	0.003	0.46
4. Walking outside on uneven terrain					
Italy (n=203)	0.34*	0.19-0.51	<0.001*	0.012	0.03
United Kingdom (n=202)	0.54	0.34-0.87	0.01	0.008	0.14
Denmark (n=201)	0.45*	0.28-0.68	<0.001*	0.015	0.17
France (n=193)	0.76	0.40-1.44	0.42	0.002	0.40
Spain (n=99)	0.61	0.26-1.36	0.16	0.014	0.01
Norway (n=62)	0.72	0.42-1.61	0.36	0.002	0.61
5. Walking inside after at least 15 minutes sitting					
Italy (n=203)	0.41	0.24-0.60	0.001	0.027	< 0.001
United Kingdom (n=202)	0.49	0.33-0.74	0.001	0.014	0.15
Denmark (n=201)	0.43*	0.29-0.65	<0.001*	0.034	< 0.001
France (n=193)	0.72	0.41-1.28	0.26	0.022	0.001
Spain (n=99)	0.29*	0.15-0.59	<0.001*	0.026	0.15
Norway (n=62)	0.16*	0.09-0.30	<0.001*	0.059	0.17
6. Walking up an incline					
Italy (n=203)	0.75	0.36-1.03	0.35	0.003	0.06
United Kingdom (n=202)	1.49	0.91-2.44	0.11	0.002	0.71
Denmark (n=201)	0.58	0.36-0.88	0.02	0.007	0.18
France (n=193)	2.26	0.89-4.53	0.02	0.010	0.17
Spain (n=99)	0.78	0.35-1.98	0.52	0.003	0.14
Norway (n=62)	1.64	0.87-3.84	0.21	0.005	0.13
7. Walking down an incline					
Italy (n=203)	0.42	0.24-0.69	0.006	0.009	0.02
United Kingdom (n=202)	1.51	0.93-2.44	0.10	0.002	0.75
Denmark (n=201)	0.64	0.41-0.97	0.05	0.007	0.09

France (n=193)	1.65	0.84-3.25	0.15	0.003	0.67
Spain (n=99)	0.81	0.35-1.78	0.57	0.004	0.12
Norway (n=62)	1.28	0.64-2.55	0.50	0.001	0.45
8. Picking up an object from the floor					
Italy (n=203)	0.26*	0.16-0.40	<0.001*	0.019	0.79
United Kingdom (n=202)	0.69	0.45-0.94	0.09	0.002	0.61
Denmark (n=201)	0.61	0.40-0.90	0.02	0.006	0.95
France (n=193)	0.52	0.29-0.96	0.04	0.013	0.05
Spain (n=99)	0.83	0.42-1.78	0.59	0.000	0.63
Norway (n=62)	1.26	0.62-2.29	0.50	0.001	0.64
9. Rising from floor					
Italy (n=203)	0.26*	0.22-0.55	<0.001*	0.025	0.03
United Kingdom (n=202)	1.18	0.57-1.25	0.41	0.004	0.10
Denmark (n=201)	1.08	0.73-1.56	0.71	0.000	0.92
France (n=193)	1.10	0.63-1.90	0.74	0.000	0.63
Spain (n=99)	0.93	0.46-2.03	0.81	0.000	0.92
Norway (n=62)	1.88	1.24-4.28	0.05	0.007	0.50
10. Rising from chair					
Italy (n=203)	2.40	1.70-4.53	0.002	0.006	0.93
United Kingdom (n=202)	0.68	0.44-1.03	0.07	0.006	0.08
Denmark (n=201)	1.21	0.80-1.83	0.38	0.004	0.06
France (n=193)	0.55	0.30-1.00	0.05	0.017	0.01
Spain (n=99)	0.94	0.49-1.91	0.85	0.011	0.001
Norway (n=62)	0.84	0.43-1.55	0.59	0.014	0.003
11. Sitting down on a chair					
Italy (n=203)	1.52	1.01-1.59	0.15	0.003	0.23
United Kingdom (n=202)	0.79	0.50-1.22	0.29	0.004	0.09
Denmark (n=201)	1.09	0.74-1.69	0.70	0.000	0.59
France (n=193)	0.55	0.29-1.06	0.08	0.008	0.08
Spain (n=99)	1.15	0.49-2.87	0.71	0.003	0.18
Norway (n=62)	1.96	0.73-3.15	0.07	0.007	0.14
12. Rising from a sofa					
Italy (n=203)	10.40*	7.28-20.68	<0.001*	0.055	0.18
United Kingdom (n=202)	1.26	0.84-1.88	0.27	0.002	0.12
Denmark (n=201)	1.26	0.86-1.87	0.25	0.001	0.90
France (n=193)	0.90	0.53-1.55	0.71	0.002	0.19
Spain (n=99)	1.39	0.70-2.67	0.30	0.001	0.96
Norway (n=62)	2.22	1.17-3.91	0.01	0.007	0.88
13. Sitting down on a sofa					
Italy (n=203)	3.75*	3.22-8.65	<0.001*	0.025	0.001
United Kingdom (n=202)	1.73	1.14-2.63	0.01	0.009	0.02
Denmark (n=201)	1.47	0.99-2.20	0.06	0.007	0.04
France (n=193)	1.74	0.96-3.14	0.07	0.004	0.28
Spain (n=99)	2.71	1.43-5.82	0.004	0.012	0.07
Norway (n=62)	1.65	0.83-2.89	0.13	0.004	0.33
14. Rising from a toilet					
Italy (n=203)	1.56	1.17-2.94	0.10	0.058#	<0.001#
United Kingdom (n=202)	1.17	0.75-1.82	0.49	0.001	0.99
Denmark (n=201)	1.57	1.09-2.52	0.04	0.006	0.22
France (n=193)	0.66	0.35-1.23	0.19	0.002	0.62
Spain (n=99)	0.79	0.35-1.57	0.51	0.001	0.51
Norway (n=62)	1.12	0.49-1.81	0.75	0.001	0.48
15. Sitting down on a toilet					
Italy (n=203)	0.88	0.58-1.48	0.64	0.024	< 0.001

United Kingdom (n=202)	0.78	0.50-1.22	0.28	0.001	0.51
Denmark (n=201)	1.16	0.78-1.87	0.51	0.001	0.72
France (n=193)	0.57	0.29-1.09	0.09	0.004	0.76
Spain (n=99)	1.07	0.51-2.78	0.85	0.006	0.18
Norway (n=62)	0.68	0.29-1.08	0.27	0.003	0.39
16. Putting on shoes					
Italy (n=203)	1.68	1.21-2.71	0.03	0.029	< 0.001
United Kingdom (n=202)	1.57	1.09-2.25	0.02	0.008	0.97
Denmark (n=201)	0.85	0.61-1.22	0.36	0.002	0.72
France (n=193)	1.46	0.89-2.40	0.13	0.019	0.006
Spain (n=99)	1.12	0.64-1.95	0.70	0.006	0.06
Norway (n=62)	0.78	0.47-1.40	0.37	0.002	0.98
17. Taking off shoes					
Italy (n=203)	1.53	0.86-1.90	0.08	0.020	0.004
United Kingdom (n=202)	1.24	0.85-1.82	0.27	0.006	0.07
Denmark (n=201)	1.15	0.80-1.64	0.46	0.002	0.55
France (n=193)	1.41	0.83-2.38	0.20	0.013	0.008
Spain (n=99)	1.43	0.82-2.69	0.24	0.003	0.90
Norway (n=62)	1.22	0.59-1.25	0.50	0.001	0.84

* Uniform DIF according the following criteria: Odds Ratio outside the interval 0.53-1.89 and $p < 0.001$

Criteria for Non-Uniform DIF are defined as a Pseudo R-square change according Nagelkerke > 0.035 and $p < 0.001$

Table 4. Spearman correlations (95% CI) between the total scores of the Animated Activity Questionnaire (AAQ), H/KOOS ADL subscale, and performance based tests in 1239 patients with hip and knee osteoarthritis.

	AAQ	H/KOOS ADL subscale	Average score performance-based tests*	Stair Climbing Test (SCT)	Timed Up and Go Test (TUG)	30 sec Chair Stand Test (CST)
AAQ	1.00	0.74 (0.71-0.76)	0.55 (0.47-0.63)†	0.68 (0.61-0.74)†	0.59 (0.50-0.66)†	0.36 (0.25-0.46)†
H/KOOS ADL subscale		1.00	0.43 (0.33-0.52)†	0.42 (0.32-0.51)†	0.37 (0.26-0.47)†	0.49 (0.39-0.58)†
Stair Climbing Test (SCT)				1.00	0.85 (0.81-0.88)†	0.34 (0.23-0.44)†
Timed Up and Go Test (TUG)					1.00	0.40 (0.29-0.49)†
30 sec Chair Stand Test (CST)						1.00

* Scores based on transformation of separate performance-based tests scores into Fisher's Z scores, calculating the average and back transformation into an average correlation score

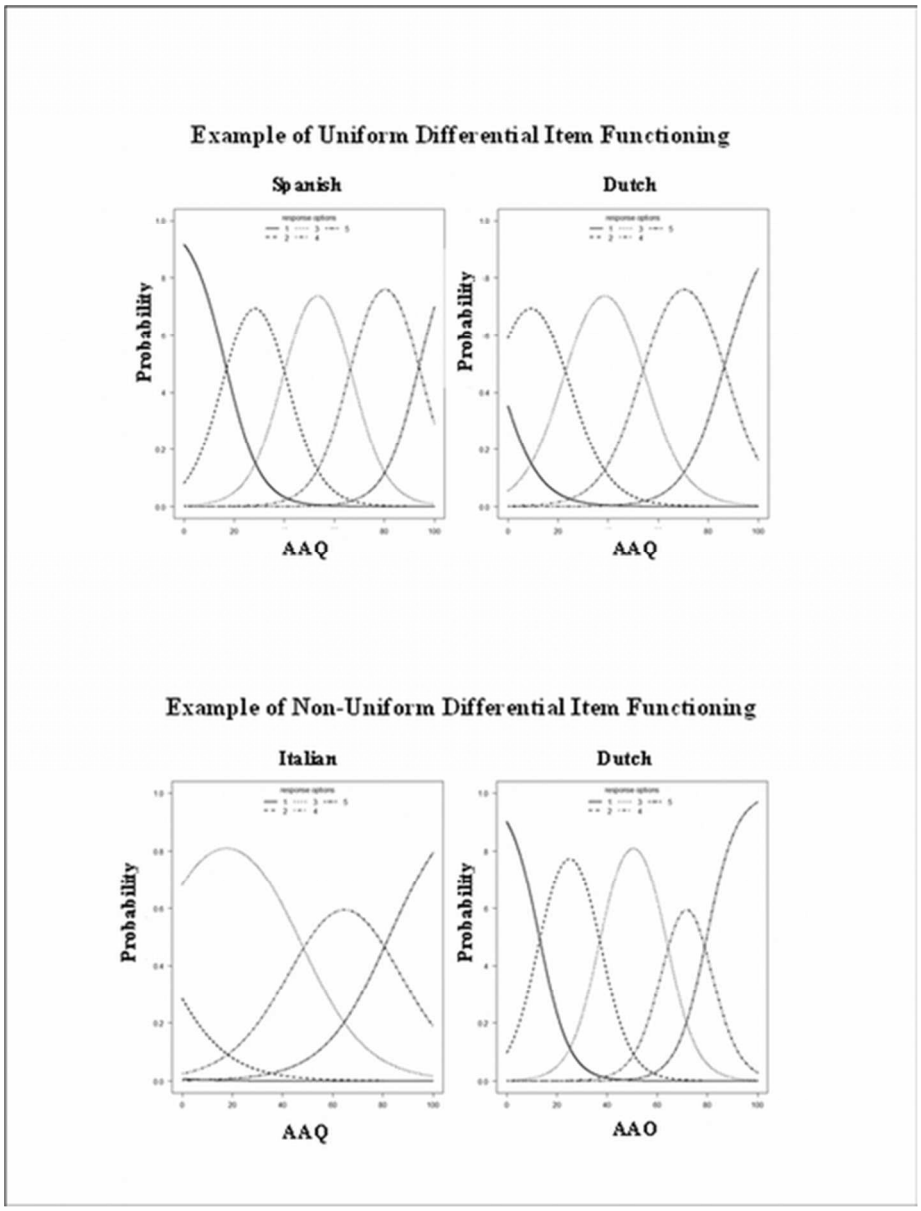
† Data analyses in a subgroup of 272 patients

Table 5 Spearman correlations (95% CI) between the total scores of the Animated Activity Questionnaire (AAQ), H/KOOS ADL subscale, and performance based tests in 7 European countries

	Correlation AAQ - H/KOOS ADL subscale		Correlation AAQ - average score performance-based tests	
Denmark	n=201	0.64 (0.55-0.72)	n=40	0.52 (0.25-0.72)
France	n=193	0.79 (0.73-0.84)	n=39	0.60 (0.35-0.77)
Italy	n=203	0.83 (0.78-0.87)	n=51	0.48 (0.24-0.67)
The Netherlands	n=279	0.79 (0.74-0.83)	n=62	0.68 (0.52-0.80)
Norway	n=62	0.78 (0.66-0.86)	*	
Spain	n=99	0.82 (0.74-0.88)	n=40	0.70 (0.50-0.83)
United Kingdom	n=202	0.85 (0.81-0.88)	n=40	0.67 (0.45-0.81)

*There was no data collected from performance-based tests in Norway

Figure 1 Examples of uniform DIF (item 5 for Spanish versus Dutch), and Non-uniform DIF (item 14 for Italian versus Dutch).



44x58mm (300 x 300 DPI)